chinaXiv:201902.00063v1

# A DYNAMIC GLOTTAL MODEL THROUGH HIGH-SPEED IMAGING[1]

**Jiangping Kong**

*Center for Chinese Linguistics, Peking University*

*Department of Chinese Language and Literature, Peking University*

*Joint Research Center for Language and Human Comlexity*

ABSTRACT

This paper is concerned with the study on a dynamic glottal model through high-speed imaging (HSI). As is well known, speech production is composed of three parts, which are speech source, speech resonance and lip radiation. Among these three parts, speech source is the most important one because it is the basis of speech. In the research of speech production, acoustical models of speech source have been well established. But the physiological speech source, that is to say, the activity of glottis is seldom researched, because the vibration of vocal folds is difficult to observe and sample. A study on glottal model was established many years ago (Kong, 2007), and in that model, the static glottis was modeled by four quarters of ellipses in three modes namely normal mode, leakage mode and open mode. The dynamic glottal control function was modeled by an approximation of multiplication of sine and exponential. The problem of the dynamic glottal model is that the control parameters can't be well explained, though the glottis can be simulated. In this study, more high-speed images were sampled, the image processing was greatly improved and the dynamic glottal control function was modeled with parameters which were significant to speech perception.

SUBJECT KEYWORDS

High-speed imaging Image, Vibration of vocal folds, Dynamic glottal model

## 1. INTRODUCTION

Speech source is very important because it is one of the three parts in speech production, which are speech source, vocal tract resonance and lip radiation. From the

chinaXiv:201902.00063v1

viewpoint of speech signal, at least three kinds of signals which are sound pressure, airflow and glottal area function can be used to model dynamic glottis. The glottal excitation, that is to say, the glottal flow and sound pressure have been studied and modeled acoustically by many researchers. But up to now, the activities of glottis are seldom studied and modeled, because the vibration of vocal folds is difficult to observe and sample. In the recent decade, more and more high-speed images in good quality were sampled through high-speed video cameras. This study focused on improving the glottal control function of the dynamic glottal model and then the application of the model was discussed.

The classical models of speech source are the one-mass model and two-mass model (Flanagan at al, 1968; Flanagan, 1969; Lucero, 1993 and 1996; Pelorson et al, 1994). These models have been established theoretically. Acoustical models based on speech sound have also been well studied. There are 7 models which are considered important. They are: 1) the acoustical model developed by Rosenberg (1971), 2) the acoustical model of Hedelin (1984), 3) the acoustical model of Fant (1979), 4) the acoustical model of Fant (1982b), 5) the acoustical model of Ananthapadmanabha (1984), 6) the acoustical model of Fant et al (1985a) and 7) the acoustical model of Ljungqvist et al (1985a). The brief definitions of these acoustical models are shown in Figure 1 to 3.

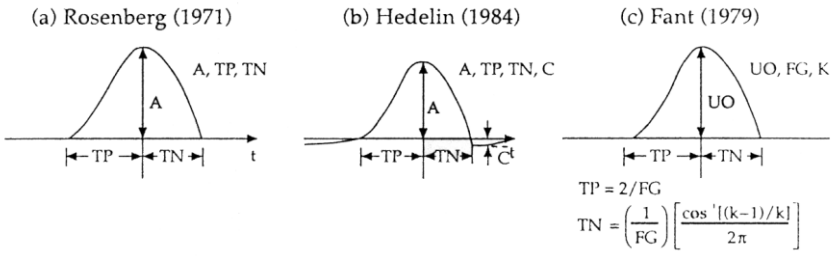



Figure 1: This figure shows the acoustical models developed by Rosenberg (a), Hedelin (b), and Fant (c).

Figure 1 shows three models that are developed in pulse of glottal flow. The model of Rosenberg was proposed in 1971. The model of Hedelin was developed in 1984. The model of Fant was set up in 1979. The brief definitions are as shown in Figure 1.

chinaXiv:201902.00063v1

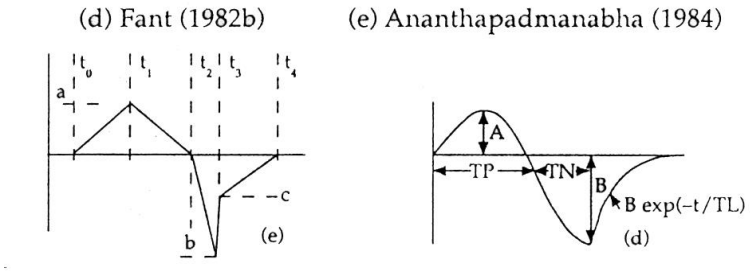



Figure 2: This figure shows models developed by Fant (d) and Ananthapadmanabha (e).

Figure 2 shows the models developed in the differential form of glottal flow, that is to say, the sound pressure. The model of Fant was established in 1982 and the model of Ananthapadmanabha was established in 1984. The brief definitions are shown in Figure 2.

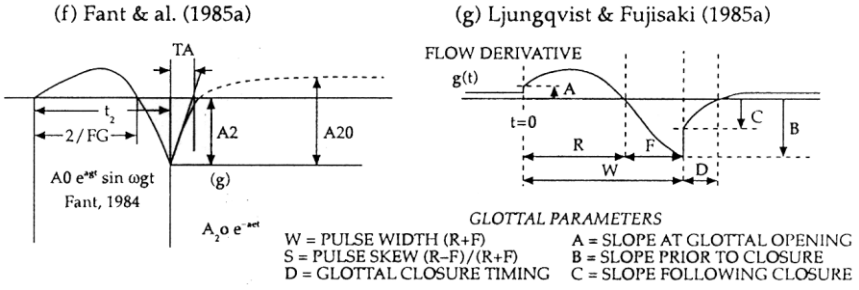



Figure 3: This figure shows the models established by Fant (f) and Ljungqvist (f).

Figure 3 shows the models established through the sound pressure. The model of Fant was established in 1985. The model of Ljungqvist et al was established in 1985. The brief definitions are shown in the figure.

Comparing these models, we can find that the three models in Figure 1 are based on glottal flow and the others in Figure 2 and 3 are based on sound pressure. It is clear that scholars first established their models in pulse of glottal flow with few parameters and then found that the models were not effective and flexible in simulating different phonation types. Then researchers tended to establish their models on the sound pressure with relatively more parameters which were more flexible and effective. Among these models, the LF-model established by Fant et al (1985a) is the most effective and flexible one.

The earlier model established by Fant used three parameters and it allowed the

change of open quotient (Fant, 1979a, 1979b, 1980). Then the LF-model reported by Fant et al in 1985 was composition of L-model (Liljecrant) and F-model (Fant) (Fant et al. 1985a).

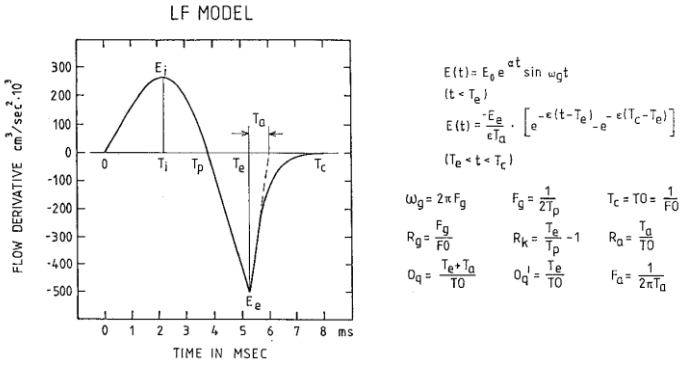



Figure 4: This figure shows the basic definition of the LF-model.

The LF-model has two phases. See Figure 4. The first phase is from 0 to 'Te', and it is created by multiplication of sine and exponential. The second phase is from 'Te' to 'Tc', which is created by the function of exponential. The formulas in Figure 4 show the relationships among the parameters. With this model, a glottal pulse of sound pressure can be specified by 4 parameters, which are 'Tp, Te, Ta and Ee'. In addition, the parameter 'Tc' is equal to 'T0', which is equal to 1/F0. Up to now, the LF-model is one of the best models in simulating different glottal source. The purpose of introducing the LF-model here is to provide a reference for a comparative study with the physiological glottal model in this paper.

There are many methods to study speech source, such as the acoustical method and physiological method. From the viewpoint of signals, speech source can be studied through acoustical signal, electroglottography (EGG) signal, air-pressure signal, high-speed image signal of vocal folds and so on. Among these signals, the signal of high-speed digital image of vocal vibration is the one which can directly and well reflect the nature of vocal vibration and explain the relationship of the movement of vocal folds and the characteristics of speech sound.

The earliest sample of vocal vibration was captured by Bell Telephone Laboratories in the 1930s, and from that time on, high speed motion pictures have been used to study the vibration of vocal folds. At present, there are some systems which can be used to study the vibration of vocal folds including the high-speed digital imaging

chinaXiv:201902.00063v1

system in the University of Tokyo, the system of kymography by Kay, the Weinberger Speedcam system, the system of the Kodak Ektapro and so on. The development of high-speed digital image systems provided a good foundation in studying the vibration of vocal folds.

## 2. GLOTTAL DETECTING

The method and procedures of this research include the sampling of high-speed digital images, image rotation and cropping, image motion compensation, image contrast adjusting and parameter extracting.

### 2.1. Sampling of high-speed images

In this study, samples of vocal fold vibration, simultaneously with the signals of EGG and speech sound were captured by high-speed imaging system produced by KAY with endoscope. The sampling rates are 2000 and 4500Hz. The samples include different phonation types, sustained vowels with low, middle and high pitches, samples of vowels with gliding voice from low to high and high to low and samples of the 4 basic tones in Mandarin.

### 2.2. Image rotation and cropping

The samples were pre-processed by rotation and cropping. The images could be rotated automatically or manually according to the concrete samples, because sometimes it was very difficult to rotate samples into the right position, especially for the disordered glottis. The samples captured by Kay system were in the size of 128×256 pixels, and they could be cropped with the size that people want to and the image in the left of Figure 5 was cropped into the size of 100×100 pixels in the right of Figure 5.

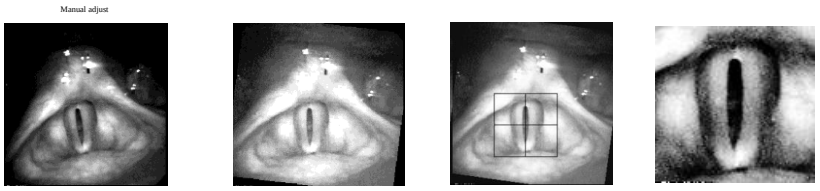



Figure 5: The first one is an original image, the second is the image which has been rotated by the algorithm of bi-cubic, the third is an image which has been windowed and the fourth is an image which has been cropped into the size of 100×100 pixels.

### 2.3. Image motion compensation (MC)

One of the adverse factors affecting the accuracy and validity of high-speed video (HSV) quantitative assessment is the motion of the endoscope's lens relative to the

chinaXiv:201902.00063v1

larynx. Endoscopic motion makes it difficult to track the dynamic characteristics of the laryngeal anatomic structures, when we divide the glottis to left and right parts, as shown in Figure 6 and 7.

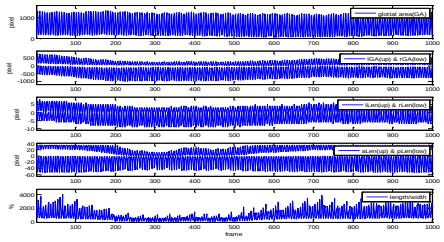


Figure 6: The glottal parameters before MC.

From up to bottom: 1) the glottal area (GA); 2) the left and right GA; 3) the left and right glottal width; 4) the anterior and posterior glottal length; 5) the ratio of glottal length to glottal width.

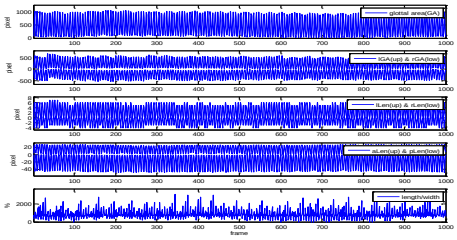


Figure 7: The glottal parameters after MC.

Therefore we have to perform "motion compensation" (MC), which means the process of detecting and removing the endoscopic motion from the HSV image. Here we mainly use the idea and method published by Dimitar D. Deliyski (2005). Figure 7 shows the result after MC. Comparing the parameters in figure 6 and 7, we can see that results after MC are very good.

2.4. Image contrast adjusting

Another adverse factor affecting HSV quantitative assessment is the contrast adjustment of the HSV image when we binarize the images to get the shape of glottis, which depends on different operators. This will influence the quantitative estimation of the glottis area and other parameters. Therefore we automatically adjust the contrast of the HSV images with the method as shown in Figure 8.

chinaXiv:201902.00063v1

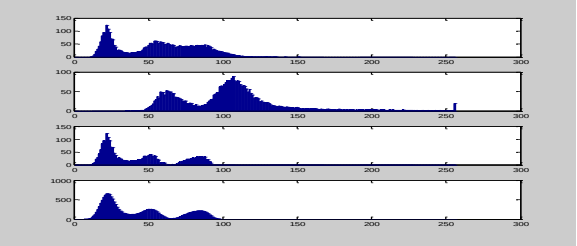


Figure8: The method of contrast adjusting.

From up to bottom: 1) The accumulated histogram when the glottis is fully opened in the first 100 frames of the video; 2) The accumulated histogram when the glottis is fully closed in the first 100 frames of the video; 3) The subtraction of them gives us the histogram of the glottis area, where is a peak in the low gray region, which reflects the gray value of the glottal region. 4) We smooth this histogram, and use the gray values at the left and right sides of the first peak to automatically adjust the contrast of all the frames in this HSV to get binarized images which show us the shape of the glottis.

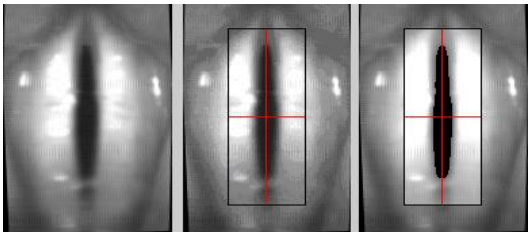


Figure 9: The left image is an original image, the middle is the image with a window and the right is the glottis detected image.

Figure 9 shows the result of contrast adjusting. The left image is an original image of modal female voice. The middle is the image with a window which is used for limiting the area of glottis. The right image shows the glottis automatically detected by the system.

2.5. Parameter extracting

After the glottis has been detected and the area of glottis has been obtained, definitions need to be given in order to extract parameters of glottis. In Figure 10, two graphs are given to help describing glottis and setting up definitions.

chinaXiv:201902.00063v1

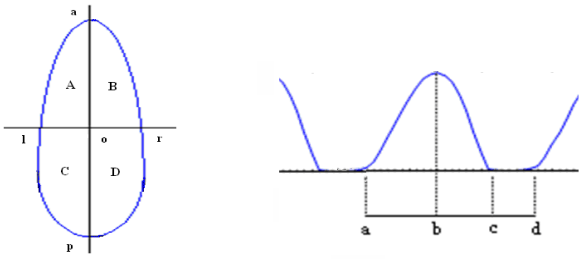



Figure 10: This figure shows the basic definition of glottis and the glottal parameters, fundamental frequency (F0), open quotient (OQ) and speed quotient (SQ).

The left in Figure 10 shows a glottis, in which 'ABCD' stands for the whole glottis, 'AB' stands for the anterior glottis, 'CD' stands for the posterior glottis, 'AC' stands for the left glottis, 'BD' stands for the right glottis. In addition, 'o' is the center of the glottis, 'lo' is the left width of glottis, 'ro' is the right width, 'ao' is the anterior length and 'po' is the posterior length. The right in Figure 10 illustrates a period of glottal area function, in which 'a' stands for glottal opening instance, 'b' stands for the local maximum of glottal area, 'c' stands for glottal close instance, and 'd' stands for the next glottal opening instance.

The following are the definitions of F0, OQ and SQ: 1) the fundamental frequency is defined as 1/'ad' (Hz); 2) The open quotient is defined as the ratio of 'ac' over 'ad'; 3) The speed quotient is defined as the ratio of 'ab' over 'bc'. See the right in Figure 10.

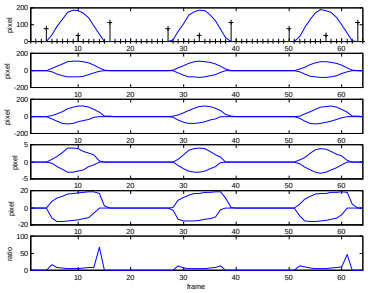



Figure 11: This figure shows the 13 parameters extracted by the system.

The Figure 11 shows the parameters extracted by our system. From up to bottom, the dynamic glottal area with glottal opening instance, glottal close instance and the local maximum; left and fright glottal areas; anterior and posterior areas; left and right

chinaXiv:201902.00063v1

chinaXiv:201902.00063v1

widths; anterior and posterior lengths; ratio of length over width. In this study, these parameters were used for dynamic glottal modeling.

## 3. MODELING ON DYNAMIC GLOTTIS

In the dynamic glottal model (Kong, 2007), the static glottis was modeled by four quarters of ellipses in three modes, namely normal mode, leakage mode and open mode. The dynamic glottal control function was modeled by an approximation of multiplication of parabola and sinusoid. The problem of the dynamic glottal model is that the control parameters can't be well explained, though the glottis can be well simulated. In this study, the static glottis was also modeled by four quarters of ellipses and the improvement of the model was focused on the dynamic glottal control function.

### 3.1. Model of static glottis

In this study, the static glottis was also modeled by four quarters of ellipses and the normal mode of static glottal model was used to explain the new dynamic glottal control function. See Figure 12.

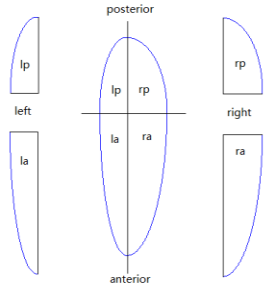



Figure12. The static glottis was modeled by four quarters of ellipses.

In Figure 12, we can see the static glottis was modeled by four quarters of ellipses, which are 'left-posterior' (lp) ellipse, 'right-posterior ellipse' (rp), 'right-anterior ellipse' (ra) and 'left-anterior ellipse' (la). The four quarters of ellipses were calculated by the two elliptical semi-major axes and two elliptical semi-minor axes respectively.

### 3.2. Glottal properties of dynamic glottis

The left and right dynamic widths and the anterior and posterior dynamic lengths in one glottal period are regarded as dynamic glottal control functions. In the dynamic model, they are the contours of the two elliptical semi-major axes and two elliptical semi-minor axes in one period, which are used to drive the dynamic glottal

model and synthesize a glottal pulse. According to the parameters extracted from the high-speed images of different phonations, the dynamic glottal control function can be classified into six basic types, which are approximations to different parts of sinusoid. See Figure 13.

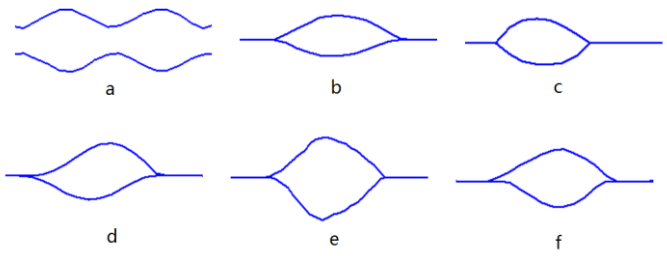



Figure13: The basic types of dynamic glottal control function

There are 6 images in Figure 13. Image ‘a’ displays two dynamic glottal control functions which are approximations to two sinusoids; image ‘b’ displays two dynamic glottal control functions which are approximations to a part of sinusoid from 300 °of the first sinusoid to 240 ° of the second sinusoid; image ‘c’ displays two dynamic glottal control functions which are approximations to a part of sinusoid from 0 °to 180 °; image ‘d’ displays two dynamic glottal control functions whose local maximum are not same, which lead to the different SQs; image ‘e’ displays two dynamic glottal control functions which are approximations to a part of sinusoid from 270 °of the first sinusoid to 180 °of the second sinusoid; image ‘f’ displays two dynamic glottal control functions whose lengths are not same which lead to the different OQs.

3.3. Modeling on dynamic glottis in open phase

In order to model the dynamic glottal control function more exactly, four parts of functions were chosen from two periods of sinusoid. See Figure 14. The angles of these two sinusoids are from 0 to 720. The chosen part of sinusoid in this study was defined by the angle of the first period of sinusoid and the angle of the second period of sinusoid respectively.

chinaXiv:201902.00063v1

chinaXiv:201902.00063v1

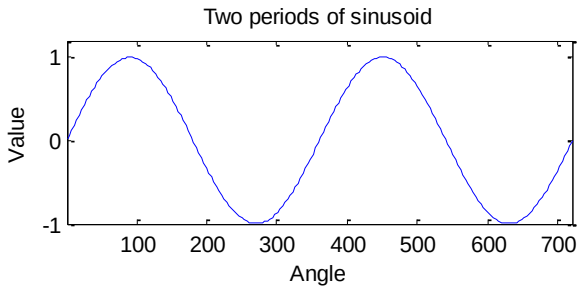



Figure14: Two periods of sinusoid are displayed in this figure. The x axis is angle, which is from 0 to 720 and value is from -1 to 1.

In these two periods of sinusoid, only the part between 270 of the first period and 270 of the second period were used for modeling the dynamic control function. According to the properties of real glottis, four typical parts of sinusoid were chosen to model the dynamic glottal control function.

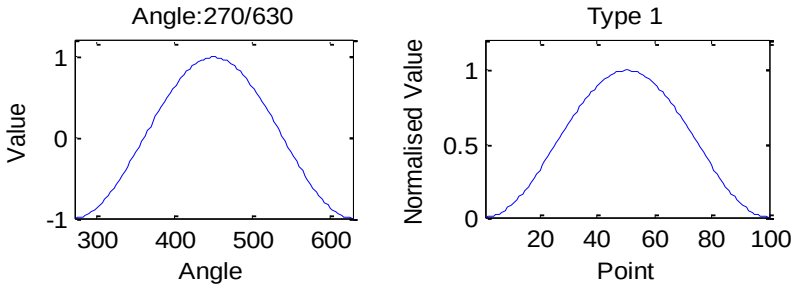
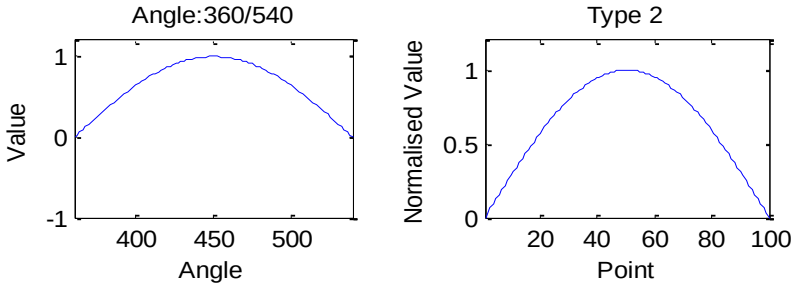



Figure15: The left image shows a part of sinusoid chosen from 270 °to 630 °of the two periods of sinusoid and the right image shows the normalized period in the left image.

In Figure 15, the left image displays a part of sinusoid from 270 °to 630 °of the two periods of sinusoid, whose value is between -1 and 1. The values are normalized from 0 to 1 for y axis and 100 points for x axis in right image. It will then be used as a part of dynamic glottal control function. It is called as ‘type 1’.



chinaXiv:201902.00063v1

Figure16: The left image shows a part of sinusoid chosen from 360 °to 540 °of the two periods of sinusoid and the right image shows the normalized period in the left image.

In Figure 16, the left image displays a part of sinusoid from 360 °to 540 °of the two periods of sinusoid, whose value is from 0 to 1. The values are normalized from 0 to 1 for y axis and 100 points for x axis in the right image. It will be used as a part of dynamic glottal control function. It is called as 'type 2'.

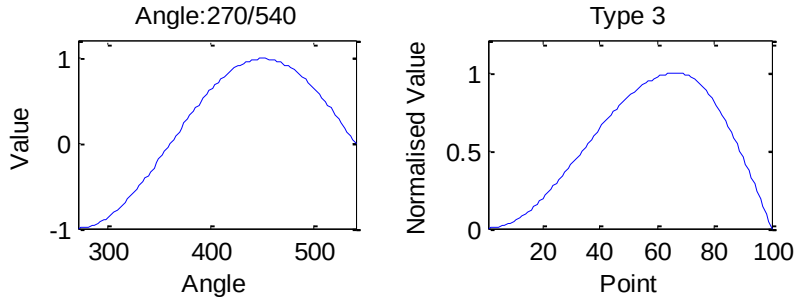



Figure17: The left image shows a part of sinusoid chosen from 270 ° to 540 ° of two periods of sinusoid and the right image shows the normalized period in the left image.

In Figure 17, the left image displays a part of sinusoid chosen from 270 ° to 540 °of the two periods of sinusoid, whose value is between -1 and 1. The right image shows the normalized period in the left image whose value is from 0 to 1 for y axis and 100 points for x axis. It will be used as a part of dynamic glottal control function. It is called as 'type 3'.

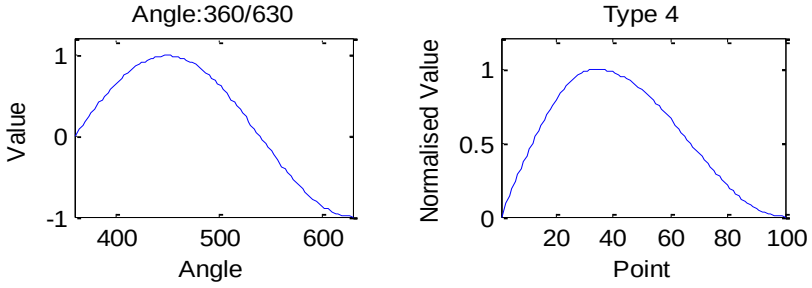



Figure18: The left image shows a part of sinusoid chosen from 360 °to 630 °of two periods of sinusoid and the right image shows the normalized period in the left image.

In Figure 18, the left image displays a part of sinusoid chosen from 360 ° to 630 °of the two periods of sinusoid, whose value is between -1 and 1. The right image shows normalized period in the left image whose value is 0 to 1 for y axis and 100 points

chinaXiv:201902.00063v1

for x axis. It will then be used as a part of dynamic glottal control function. It is called as 'type 4'.

3.4. Modeling on dynamic glottal control function

Based on the four types above, the parameters of F0, OQ and SQ were added to produce a whole dynamic glottal control function. In order to explain easily, the angles chosen from the two periods of sinusoid will be set and explained separately. See Figure 19.

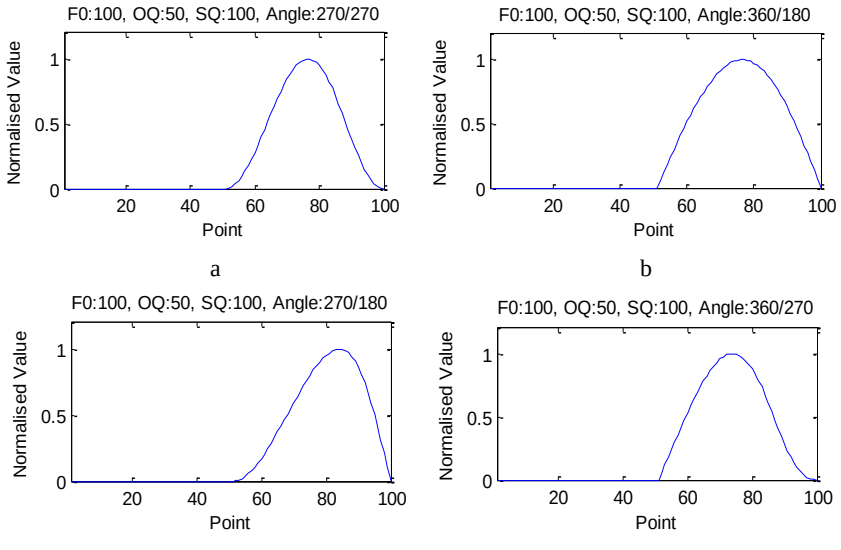



Figure 19: In this figure, there are four plots which show the four typical types of dynamic glottal function. Plot 'a' shows that of type1, plot 'b' shows that of type 2, plot 'c' shows that of type 3, and plot 'd' shows that of type 4.

In Figure 19, plot 'a' displays the dynamic glottal control function of type 1, whose F0 is 100 Hz (the sampling rate is 10k), OQ is 50% and SQ is 100% and the pulse in open phase is chosen from 270 ° of the first period of sinusoid to 270 ° of the second period of sinusoid; plot 'b' displays the dynamic glottal control function of type 2, whose F0 is 100 Hz, OQ is 50% and SQ is 100% and the pulse in open phase is chosen from 360 ° of the first period of sinusoid to 180 ° of the second period of sinusoid; plot 'c' displays the dynamic glottal control function of type 3, whose F0 is 100 Hz, OQ is 50% and SQ is 100% and the pulse in open phase is chosen from 270 ° of the first period of sinusoid to 180 ° of the second period of sinusoid; plot 'd' displays the dynamic glottal

chinaXiv:201902.00063v1

control function of type 4, whose F0 is 100 Hz, OQ is 50% and SQ is 100% and the pulse in open phase is chosen from 360 ° of the first period of sinusoid to 270 ° of the second period of sinusoid.

## 4. MODAL VOICE SYNTHESIS

According to the model of static glottis and the 4 types of glottal control functions, different pulses of phonation types can be synthesized by parameters of 4 F0s, 4 OQs, 4 SQs, 2 angles of sinusoid, 2 widths of glottis and 2 lengths of glottis.

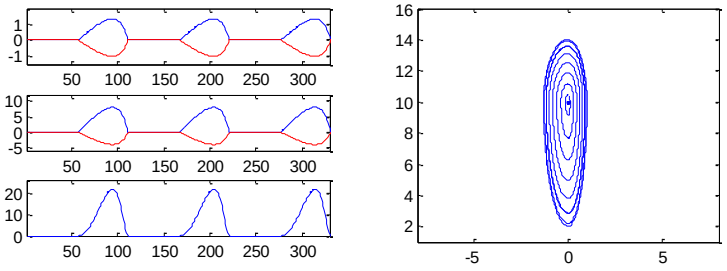
### 4.1 Modal voice synthesis

Modal voice is a kind of phonation type which is most popular used in spoken language. Usually the F0 is around 70 and 300, the SQ is around 100 and 300% and the OQ is around 50 and 60%. See Table 1.

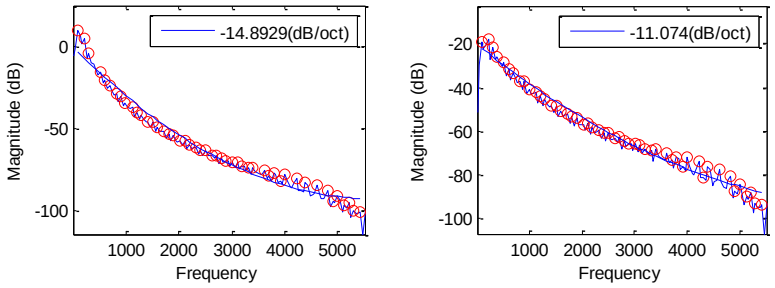
Table1.The synthesis parameters of a modal voice

| Semi axis | F0 | OQ | SQ | Angle 1 | Angle 2 | width/length |
|---|---|---|---|---|---|---|
| Left | 100 | 50 | 300 | 360 | 180 | 1.3 |
| Right | 100 | 50 | 300 | 360 | 180 | 1 |
| Anterior | 100 | 50 | 300 | 360 | 180 | 8 |
| Posterior | 100 | 50 | 300 | 360 | 180 | 4 |

In Table 1, the basic parameters for voice synthesis are listed. ‘Left’ stands for left width; ‘Right’ stands for right width; ‘Anterior’ stands for anterior length and ‘Posterior’ stands for posterior length. For synthesizing a modal voice, the parameters of 100 (Hz), 50% , 300%, 360 ° and 180 ° for F0, OQ, SQ, angle 1 and angle 2 are set. The left width is 1.3mm, the right width is 1mm, the anterior length is 8mm and the posterior length is 4mm. The dynamic parameters and the synthesized dynamic glottal area are displayed plot ‘a’ in Figure 20.




a b




c d

Figure20: Parameters, synthesized glottal area and the spectrums.

There are four plots in Figure 20. Plot ‘a’ displays the left width, right width, anterior length, posterior length and glottal area from top to bottom. Plot ‘b’ displays the dynamic glottises overlapped together. Plot ‘c’ displays the spectrum of glottal area function, which is -14.8929 dB/oct. Plot ‘d’ displays the spectrum of glottal area in differential form, which is -11.074dB/oct.

4.2 Modal voice synthesis with different SQ

In the vibration of human vocal folds, the left and right vocal folds often do not abduct and adduct at the same time in one vocal period. The basic parameters for synthesizing such a voice are listed in Table 2.

Table 2. The basic parameters of a modal voice with different SQs

| Semi axis | F0 | OQ | SQ | Angle 1 | Angle 2 | width/length |
|---|---|---|---|---|---|---|
| Left | 100 | 50 | 300 | 360 | 180 | 1.5 |
| Right | 100 | 50 | 75 | 360 | 180 | 1.5 |
| Anterior | 100 | 50 | 300 | 360 | 180 | 8 |
| Posterior | 100 | 50 | 75 | 360 | 180 | 4 |

In Table 2, the basic parameters of a modal voice are listed. The SQs of left and right widths and anterior and posterior lengths are not the same. The SQs for the dynamic glottal control function of left widths and anterior length are 300% and the SQs for the dynamic glottal control function of right width and posterior lengths are 75%. The angles for choosing the parts from the sinusoids are 360 ° and 180 °. The widths of left and right glottis are 1.5 mm and the lengths of anterior and posterior glottis are 8 mm and 4 mm respectively.

chinaXiv:201902.00063v1

chinaXiv:201902.00063v1

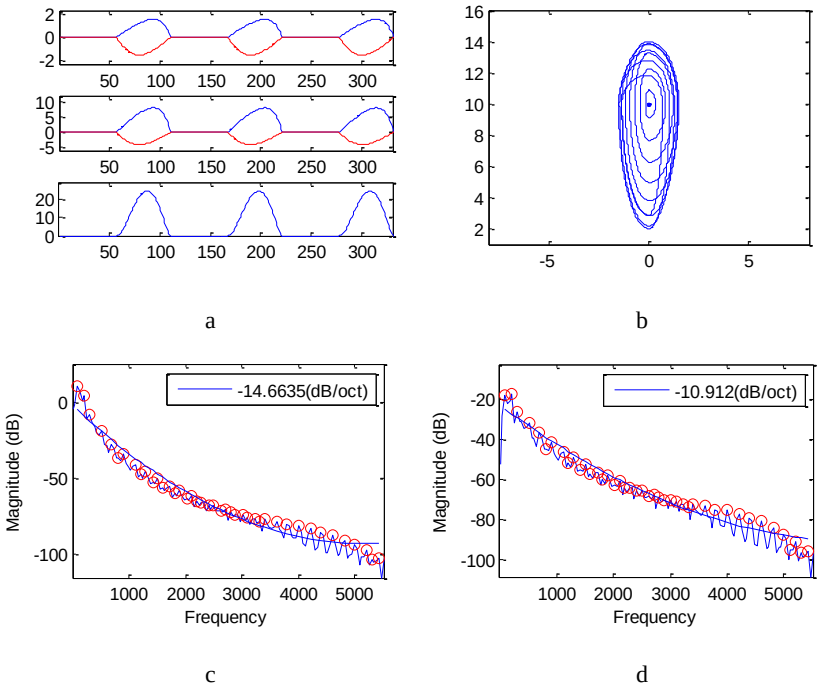



Figure21: This figure shows the synthesized parameters, the dynamic glottal areas, the two spectrums of glottal area and the glottal area in the differential form.

There are four plots in Figure 21. Plot ‘a’ displays the left width, right width, anterior length, posterior length and glottal area from top to bottom. Plot ‘b’ displays the dynamic glottis overlapped together. Plot ‘c’ displays the spectrum of glottal area function, which is -14.6635 dB/oct. Plot ‘d’ displays the spectrum of glottal area in differential form, which is -10.912 dB/oct. From the parameters and synthesized glottal areas, we can find that the acoustical models talked above can’t synthesized such glottal pulses, which is more flexible than those synthesized by the acoustical models.

4.3 Falsetto synthesis

Falsetto is a kind of voice with a very high pitch. Falsetto is not usually used in normal spoken language but often used in oral performance such as singing and opera. In Chinese oral cultures, such as Kunqu and Peking opera, falsetto is often used by Dan (young female) players. In Table 3, the parameters for synthesizing a falsetto are listed.

Table3. Falsetto synthesis parameters

| Semi axis | F0 | OQ | SQ | Angle 1 | Angle 2 | width/length |
|---|---|---|---|---|---|---|
| Left | 400 | 100 | 100 | 315 ° | 225 ° | 0.7 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Right | 400 | 100 | 100 | 315 ° | 225 ° | 0.7 |
| Anterior | 400 | 100 | 100 | 315 ° | 225 ° | 7 |
| Posterior | 400 | 100 | 100 | 315 ° | 225 ° | 5 |

In Table 3, we can see that the F0 is 400 Hz, which is very high for a male speaker. The OQ is 100% which is the largest OQ in the vibration of vocal folds and the SQ is 100%, which is very small. The small SQ indicts the small power in high frequency. The part of sinusoid was chosen from 315 ° of the first period of sinusoid to 225 ° of the second sinusoid. The widths of left and right glottis are 0.7 mm and the lengths of anterior and posterior glottis are 7 mm and 5 mm which mean that the glottis is very narrow and long. The synthesized daynmic control function, the glottal area function, the dynamic glottis, and the spectrums of glottal area function and the differential form of glottal area function are showed in Figure 22.

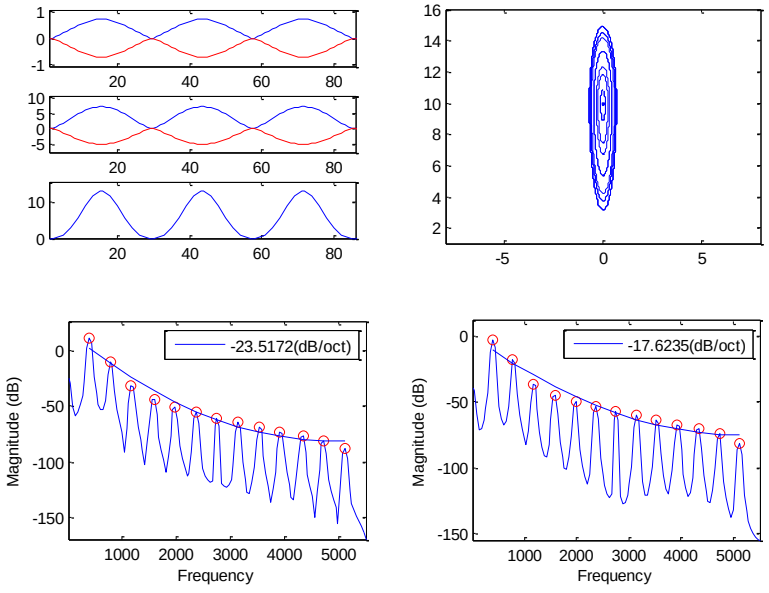



Figure 22: This figure displays the synthesized daynmic glottal control function, the glottal area function, the dynamic glottis, and the spectrum of glottal area and the differential form of glottal area.

In Figure 22, we can see the synthesized glottal area function looks like sinusoid very much and the glottis looks narrow and long. The spectrum of glottal area function is -23.5172 dB/oct, which means that the power in high frequency is small, and

chinaXiv:201902.00063v1

chinaXiv:201902.00063v1

the spectrum of differential glottal area function is -17.6235 dB/oct which means that the power is still small in the speech production of human voice.

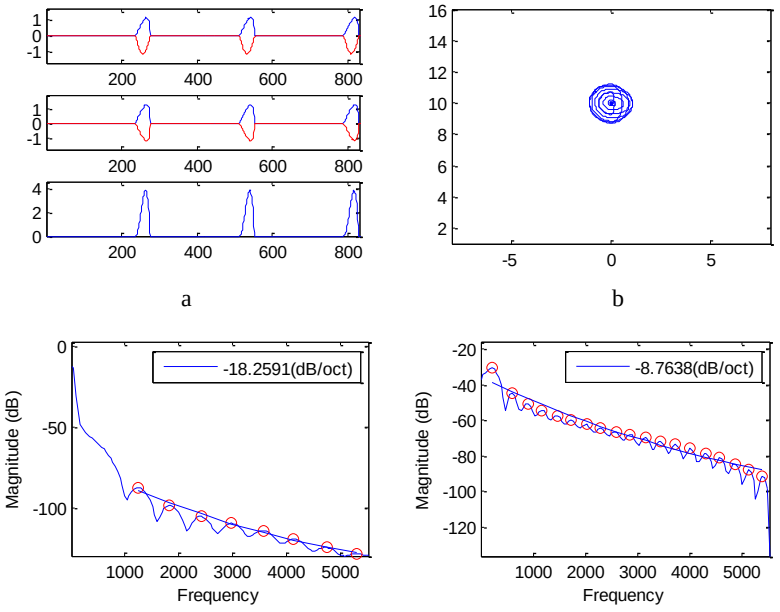
4.4 Vocal fry synthesis

Vocal fry is a kind of phonation with very low ptich and large power in high frquency. Vocal fry sometimes appears in the middle of low tone (tone 3) in Mandarin. This phonation is very close to that of creaky voice which usually has irragular periods. The basic parameters for synthesizing a fry voice are listed in Table 4.

Table4. Vocal fry synthesis parameters

| Semi axis | F0 | OQ | SQ | Angle 1 | Angle 2 | width/length |
|---|---|---|---|---|---|---|
| Left | 40 | 15 | 300 | 360 | 180 | 1.1 |
| Right | 40 | 15 | 100 | 360 | 180 | 1.1 |
| Anterior | 40 | 15 | 300 | 360 | 180 | 1.3 |
| Posterior | 40 | 15 | 300 | 360 | 180 | 1.2 |

In talbe 4, the F0 is 40 Hz which is very low for both male and femanle speakers. The OQ is 15%, which is very small and the SQ is 300% for left width, anterior length and posterior length, and 100% for right width of glottis. The angle 1 is 360 ° of the first period of sinusoid and the angle 2 is 180 ° of the second period of sinusoid. The left and rigth widths are 1.1mm and the left and right lengths are 1.3 mm and 1.2 mm. The synthesized parameters and spectrums are showed in Figure 23.



c d

Figure23: This figure displays the synthesized daynmic glottal control function, the glottal area function, the dynamic glottis, and the spectrum of glottal area and the differential form of glottal area of a vocal fry.

In Figure 23, we can see the synthesized glottal area function looks like a saw waveform, and the glottis looks round and small. The spectrum of glottal area function is -18.2591 dB/oct, which indicts that the power in high frequency is not high, and the spectrum of differential glottal area function is -8.7638 dB/oct, which means that the power is very large in the speech production of human voice.

4.5 Synthesis of a Diplophonia

Diplophoia is a kind of voice with different fundamental frequencies of left and right vocal folds. It is not a normal voice in human speech but often appears in disordered voice. Sometimes people would sound diplophonia for singing perpformence. The basic parameters for synthesizing a diplophnia are listed in Table 5. The synthesized parameters and spectrums are showed in Figure 24.

Table5. Basic parameters for synthesizing a diplophonia

| Semi axis | F0 | OQ | SQ | Angle 1 | Angle 2 | width/length |
|---|---|---|---|---|---|---|
| Left | 200 | 100 | 100 | 270 | 270 | 1.5 |
| Right | 180 | 100 | 100 | 270 | 270 | 1.5 |
| Anterior | 200 | 100 | 100 | 270 | 270 | 6 |
| Posterior | 180 | 100 | 100 | 270 | 270 | 4 |

In talbe 5, the F0 of left glottis is 200 Hz and the F0 of right glottis is 180 Hz. These two F0s are not same whose difference is 20 Hz. The F0 of anterior glottis is 200 Hz and the F0 of posterior glottis is 180 Hz. These two F0s also have 20 Hz difference. The OQ is 100%, which are very large and SQ is 100%, which is very samll. The angle 1 is 360 ° of the first period of sinusoid and the angle 2 is 180 ° of the second period of sinusoid. The left and rigth widths are 1.5 mm and the anterior and posterior lengths are 6 mm and 4mm. The synthesized parameters and spectrums are showed in Figure 24.

chinaXiv:201902.00063v1

chinaXiv:201902.00063v1

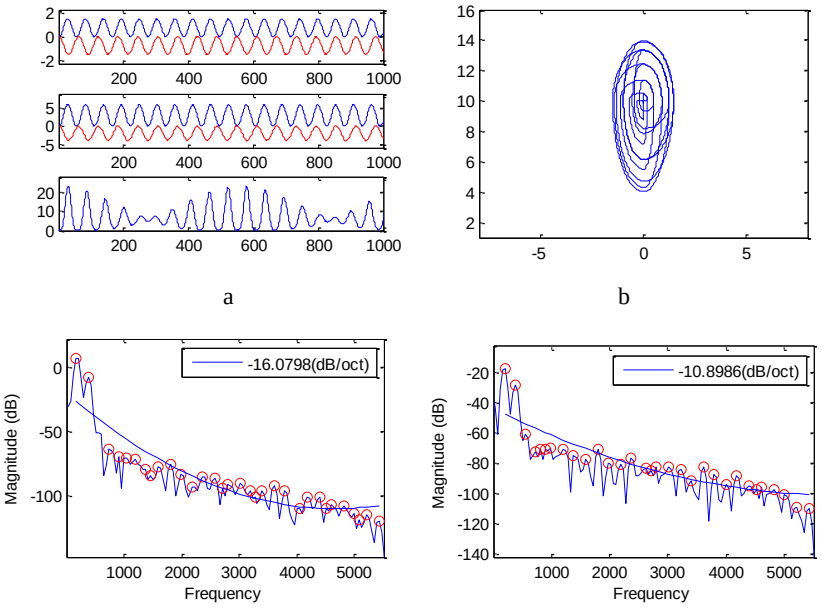



Figure 24: This figure displays the synthesized daynmic glottal control function, the glottal area function, the dynamic glottis, and the spectrum of glottal area and the differential form of glottal area of a Diplophonia.

In Figure 24, the daynmic glottal control functions, the dynamic glottis and the spetrums of glottal area functions are displayed plot ‘a’, from which we can see the F0s of left and right widths of glottis are not same and a super period which covers around 10 vocal periods. In plot ‘b’, we can see that the glottis is not symmetry in vibration. The spectrum of glottal area function is -16.0798 db/oct and the spectrum of the differential glottal area function is -10.8986 db/oct which indicts that the power is not small in the high frequency, though the SQs are small. This sample can’t be synthesized by any acoustical models, which tells us that the glottal model is in deeper level which is more effective and flexible in source production and the acoustical model is at the surface level.

## 5. CONCLUDING REMARKS

As is well known, voice models can be studied from the viewpoints of speech science, phonetics, speech engineering and so on, and established for different purposes, such acoustical model which may be used in speech synthesis, physiological model

chinaXiv:201902.00063v1

which may be used to imitate the physiological activity of vocal folds for medical purpose, phonetic model which may be used to study the linguistic significance of phonation types and singing model which may be used in synthesizing special songs or in singing teaching. In this research, the dynamic glottal control functions were studied and improved on the basis of physiological model first established by Kong (2001) and published in 2007. The improvement in this study includes three kinds of basic parameters which are: 1) the basic parameters of F0, OQ and SQ which are the most popular parameters in the study of speech science and phonetics and whose properties have close relationship with human perception and linguistic significance; 2) the basic parameters of 4 types of dynamic glottal control functions, which are chosen from parts of a sinusoid defined by two angles; 3) the basic parameters of glottal size which are widths of left and right glottis and lengths of anterior and posterior glottis. With this improved physiological model, different kinds of phonation types can be simulated and further studied for many purposes and in many fields.

Although the model can now be easily used to synthesize many different phonation types, many things can still be further studied and improved. First of all, since the sampling rate of high-speed image system is not high enough, the parameters extracted from the voice sample with high pitch are not very accurate, especially for the parameter of SQ and F0, because the position of peak in one period of glottal area function is not stable. Secondly, the motion compensation and automatic contrast adjusting of the image processing system can also be improved. Thirdly, for the sake of sampling rate, it is still difficult to study spectrum of glottal area function and the relationship between glottal area function and the other signals. We believe that along with the development of high-speed image system, good samples with high sampling rate, maybe 3D samples, can be captured for modeling a 3D dynamic glottal model to simulate the vibration of vocal folds and synthesize different phonations with more accurate dynamic parameters.

## NOTES

1. This research is funded by the National Natural Sciences Foundation of China (No: 61073085). We would also like to give our thanks to prof. Edwin Yiu in the University of Hong Kong for the high-speed image sampling, all the subjects and Wang Gaowu for program improvement of the high-speed image system.

## REFERENCES

chinaXiv:201902.00063v1

KONG Jiangping, 2007. Laryngeal Dynamics and Physiological Models, Peking University Press.

FLANAGAN J.L.and Landgraf L.L. (1968). Self-oscillating source for vocal-tract synthesizers. IEEE Trans.16, March 1968, 57-64.

LUCERO J. C. 1993. Dynamics of the two-mass model of the vocal folds: equilibria, bifurcations, and oscillation region. J. Acoust. Soc. Am. 94(6), Decmber.

LUCERO J. C. 1996. Chest- and falsetto-like oscillations in a two-mass model of the vocal folds.J. Acoust. Soc. Am. 100 (5), November.

PELORSON X., Hirochberg A., Hassel van R.R., Wijnands A.P.J., Auregan Y. 1994. Theoretical and experimental study of quasi-steady flow separation within the glottis during phonation. Application to a modified two-mass model a)", J. Acoust. Soc. Am. 96(6), December.

ROSENBERG A.E. (1971). Effect of glottal pulse shape on the quality of natural vowels. Journal of the Acoustical Society of America, 49,583-98.

HEDELIN P. (1984). A glottal LPC-vocoder. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1.6.1-1.6.4. San Diego.

ANANTHAPADMANABHA T.V. (1984). Acoustic analysis of voice source dynamics. Speech Transmission Laboratory – Quarterly Progress and Status Report, 2-3, 1-24. Roya Institute of Technology, Stockholm.

LJUNGQVIST M. and Fujisaki H. (1985). A comparative study of glottal waveform models. Technical Report of the Institue of Electronics and Commmunications Engineers. Japan, EA85-58, 23-9.

FANT G. (1979a). Glottal source and excitation analysis. STL-QPSR, No. 1, pp. 85-107.

_____.1979b. Voice source analysis – a progress report. STL-QPSR, Nos. 3-7, pp. 31-54.

_____.1980. Voice source dynamics. STL-QPSR, Nos. 2-3, pp. 17-37.

_____.1982b. The voice source, acoustic modeling. STL-QPSR, No. 4, pp. 28-48.

_____.Liljencrants J. and Lin Q. (1985a). A four parameter model of glottal flow. STL-QPSR, No. 4, 1985, pp 1-13.

_____.and Lin Q. (1988). Frequency domain interpretation and derivation of glottal flow parameters. STL-QPSR, Nos. 2-3, pp. 1-21.

DIMITAR D. Deliyski, (2005). "Endoscope Motion Compensation for Laryngeal High-Speed Video endoscopy", Journal of Voice, Vol. 19, No. 3, pp. 485-496

chinaXiv:201902.00063v1

KONG Jiangping, (2001), “Study on Dynamic Glottis: though High-Speed Digital Imaging”, (in English), Ph.D. dissertation, City University of Hong Kong, Hong Kong, China

# 基於高速數位成像的動態聲門模型

孔江平
北京大學中國語言文學系
中國語言學研究中心
語言與人類複雜系統聯合研究中心

提要

本文利用高速數位成像技術對動態聲門模型進行了研究。眾所周知，言語產生包括嗓音聲源、聲道共鳴和唇輻射三個方面，其中嗓音聲源尤其重要，因為嗓音聲源是言語產生的基礎。在言語產生的研究中，聲學模型已經有了很深入的研究，但由於聲帶振動難於觀察和採集樣本，嗓音的生理模型研究的很少。多年前作者建立了一個動態聲門模型（Kong，2007），在此模型中，靜態聲門是用四個四分之一橢圓來建模的，並有正常、漏氣和敞開四種模式。模型的動態聲門控制函數是通過正弦和拋物線的乘積來建模。雖然這種方式有效，但合成嗓音的參數解釋性較差。在本項研究中，採集了更多更高品質的聲帶振動高速數位成像樣本和大大改進了數位影像處理的技術，最終模型的動態聲門控制函數所用的參數對嗓音聲源的感知具有很好的解釋性。

關鍵字
高速數位成像　聲帶振動　動態聲門模型